INDIVIDUAL DIFFERENCES IN THE EFFECTS OF RETRIEVAL

FROM LONG-TERM MEMORY

by

GENE A. BREWER JR

(Under the Direction of Nash Unsworth)

ABSTRACT

The current study examined individual differences in the effects of retrieval from long-term memory (i.e., the testing effect). The effects of testing memory make tested information more accessible for future retrieval attempts. Despite the broad applied ramifications of such a potent memorization technique there is a paucity of research tailored toward explaining variability in the effect. Multiple measures of working memory capacity, attention control, episodic memory, and general-fluid intelligence were collected in addition to performance in a standard paired-associate testing task. A testing effect was observed and there was a great deal of individual variability in the magnitude of the effect. This variability was best accounted for by memory and intelligence constructs. Furthermore, the pattern of results is consistent with the notion that students with poor memory abilities and substandard intelligence benefit more so from testing memory than high ability students.

INDEX WORDS: TESTING EFFECT, WORKING MEMORY, EPISODIC MEMORY, ATTENTION CONTROL, INTELLIGENCE, PAIRED ASSOCIATES, CUED RECALL

INDIVIDUAL DIFFERENCES IN THE EFFECTS OF RETRIEVAL FROM LONG-TERM MEMORY

by

GENE A. BREWER JR

B.A., University of Georgia, 2003

M.S., University of Georgia, 2005

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2010

INDIVIDUAL DIFFERENCES IN THE EFFECTS OF RETRIEVAL FROM LONG-TERM MEMORY

by

GENE A. BREWER JR

Major Professor: Nash Unsworth

Committee: Richard Marsh

Brett Clementz

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2010

DEDICATION

Any individual is lucky if they have a loving father to support them throughout their life. I have been quite fortunate in that I have had no less than four "Dads" who have made consequential, independent, and lasting contributions to my development as a scholar and the man that I am today. First and foremost, the efforts of my father Gene A. Brewer Sr., Ph.D. have fortified me and been a consistent presence at the highest zenith, the lowest nadir, and every step in between. Father, without your unconditional love, support, encouragement, wisdom, and sage-like advice there is little doubt that I would be much less than I am today. Second, Clifford S. Proffer has been a source of strength, integrity, and inspiration for me since I met him years ago at my high school baseball game. Cliff, although there may be more differences than similarities between us, you can forever trust that I will spend a great deal of my time trying to be more like you every day. Third, Richard L. Marsh, Ph.D. can single handedly be credited with teaching me how to think like an academic. Rich, the generous appropriation of your time has resulted in me becoming a true experimental psychologist and scholar. Fourth, it is my opinion that Nash Unsworth, Ph.D. has served as a shining example of what all young academics should strive for in their careers. Nash, your appetite for knowledge, ability to synthesize information, and unrelenting productivity makes you a prototypical example of a true scholar. I can testify with unwavering certainty that working with you has changed my approach to studying human memory and the way that I conduct research in general. Dads, all of you have been there for me in my times of need and for that I am thankful and will always strive to make you proud. So, for all of your efforts, this dedication is for you! You can never understand how grateful that I am to be fortuitous enough to write it.

ACKOWLEDGEMENTS

There are many people that have supported me while I completed my graduate career and it would take another dissertation to thank them all. I owe you all a great deal of thanks. Specifically, I would like to acknowledge my mother for all of the support that she has given me over the years. Also, I would like to acknowledge Kristel for helping me in almost every facet of my life over the past two years. You are the best! I would also like to thank my brother and sisters for allowing me to be myself and loving me no matter what. All of you have given me so much over the years and I would not be here if it were not for you.

Academically, I have been in the presence of scholars ever since matriculating at the University of Georgia. Jaxk, thank you for helping me navigate graduate school and being a great teammate in every regard. To Brett, thank you for always giving me scholarly advice and providing me with alternative points of view. To my fellow lab mates - Greg, Justin, Arlo, and Thad - I have enjoyed all of our conversations including the ones where, looking back, it is clear that I was horribly misguided in my thinking and ignorant in my assumptions. I consider myself lucky to have been in the company of scholars over the past 7 years.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

INTRODUCTION

Individual Differences in the Effects of Retrieval from Long-Term Memory

> "The relationship between test scores and school performance seems to be ubiquitous. Wherever it has been studied, children with high scores on tests of intelligence tend to learn more of what is taught in school than their lower scoring peers. There may be styles of teaching and methods of instruction that will decrease this correlation, but none that consistently eliminates it has yet been found…" - Neisser et al. (1996, p. 82)

The act of retrieving information from memory reinforces that information, and related information, thereby rendering it more accessible for later retrieval attempts (Abbott, 1909; Bjork, 1975). The effect of retrieval on subsequent memory performance is referred to as the testing effect. Recently, there has been a surge of interest and empirical research applying cognitive principles to understanding the testing effect (for a review see Roediger & Karpicke, 2006). Results from these studies indicate that the testing effect extends across various memory tasks, measures of memory, testing schedules, and materials. The principal reason for the recent interest in investigating the testing effect is the potential benefits of implementing repeated testing in the classroom (McDaniel, Roediger, & McDermott, 2007). This research has greatly enhanced our theoretical and empirical understanding of the benefits of testing memory, but there are still many unresolved issues. For instance, there is limited individual differences research investigating the testing effect. That is, the benefits of retrieving information from memory may not extend to all people in the same manner and it is important to understand what individual difference variables are related to the effects of testing. The primary goal of the current report is to provide one of the first individual differences analysis of the testing effect.

The Testing Effect

In standard testing-effect paradigms, a set of to-be-remembered material is encoded and then subsequently retrieved. After the initial testing period participants engage in some other activity or delay before having their memory for that same information probed again later in the future. Memory for intermediately tested information is more immune to forgetting and is also more accessible for future retrieval attempts (Karpicke & Roediger, 2006). The benefits of testing extend beyond mere re-presentation (restudy) and are generally apparent after some significant delay (e.g., greater than an hour: Carrier & Pashler, 1992). Thus, retrieval makes memories for tested materials more durable than memories for restudied material even though more test-relevant information is processed through representation than retrieval. Several methodologies have been used to examine the beneficial effects of testing on subsequent memory including free recall, recognition, and cued recall. The results from a multitude of studies implementing different testing procedures have supported the hypothesis that one causal mechanism underlying the testing effect is controlled and effortful retrieval from long-term memory (Carpenter, in press; Carpenter & Delosh, 2006; Chan & McDermott, 2007; Kang, McDermott, & Roediger, 2007; Glover, 1989). Given the reliability of the testing effect across a variety of paradigms (see Roediger, Agarwal, Kang, & Marsh, in press for a review), the current study will focus specifically on cued-recall paradigms.

Cued recall, or paired-associate learning, is a long-standing technique for studying associative memory (Calkins, 1894). In a typical cued-recall task participants study A-B pairs and after some variable retention interval they are re-presented with the cue (A) and they try to reproduce the target (B). With respect to the testing effect, the cued-recall paradigm has been fruitful for exploring the cognitive processes underlying retrieval benefits (Carpenter & Delosh,

2005; Carpenter, Pashler, & Vul, 2006; Carrier & Pashler, 1992; Cull, 2000; Izawa, 1969). Carpenter et al. (2006) reported that the association between A-B pairs was strengthened by testing as compared with restudy. In the intermediate testing phase, participants were always given the cue (A) and they tried to retrieve the target (B). On the final test, participants were either given the original cue (A) or the original target (B) and they were asked to produce the other member of the pair. Across conditions, participants consistently recalled the pair member with higher probability when it had been previously tested roughly 18 to 48 hours earlier. This result extended prior theorizing in the testing literature by suggesting that retrieval strengthens associative information in a bidirectional manner. The strengthening of associative information reported by Carpenter and colleagues is also consistent with recent reports that recollective details are enhanced by intermediate free recall testing (Chan & McDermott, 2007).

Early treatments of the testing effect attempted to describe the potentiating effects of retrieval in terms of Estes' (1959) stimulus-sampling model (Izawa, 1967, 1969; see also Murdock, 1974 for a brief review). Memory researchers who are revisiting the testing effect have proposed several alternative theoretical explanations for the beneficial effects of retrieval on associative memory (Roediger & Karpicke, 2006). One early theoretical proposition claimed that testing provided students with additional study time, but this claim was quickly invalidated (e.g., Carrier & Pashler, 1992). Another hypothesized explanation suggests that transfer-appropriate processing is an important mechanism underlying the testing effect (McDaniel, 2007). Although the notion of transfer-appropriate processing seems amenable to explaining the testing effect, two recent articles suggest that intermediate free recall testing generally provides stronger benefits to final tests even when the final tests do not match the intermediate tests (Carpenter & DeLosh, 2006; Kang et al., 2007). As mentioned previously, effortful processing

during intermediate testing appears to be the causal mechanism underlying the testing effect (Bjork, 1975). A corollary to this hypothesis is that effortful retrieval increases the number of routes (viz. associations) converging on the critical item leading to enhanced subsequent memory for that item (Bjork, 1988; McDaniel & Masson, 1985). Of course, depth of processing is not entirely inconsistent with the notion of transfer-appropriate processing. Nevertheless, disentangling these explanations is an important endeavor.

An extension of the retrieval-effort hypothesis has been proposed by Carpenter and Delosh (2006). Specifically, they argued that the retrieval processes that are operating during an initial test are responsible for improved memory later on. Carpenter (in press) investigated the retrieval-effort hypothesis by manipulating the pre-experimental associative strength of the A-B pairs in a cued-recall paradigm. Briefly, participants benefited more from intermediate tests on weakly associated A-B pairs than from testing on strongly associated A-B pairs. Theoretically, Carpenter proposed that the easily accessible target information in the strongly associated A-B pairs reduced the degree of effortful retrieval necessary for successful recall and thus diminished the positive effects of testing on subsequent memory performance (see Pyc & Rawson, 2009 for similar findings). These theoretical explanations are typically evoked to describe direct effects of retrieving information from memory, but there are also indirect effects of intermediate testing (Roediger & Karpicke, 2006).

Indirect (i.e., mediated) effects of retrieval are found when students capitalize on previous retrieval attempts to fine tune their metacognitive monitoring and control processes (Nelson & Narens, 1990). Beyond strengthening associations, an additional benefit of testing memory is found when students realize what information they can and cannot retrieve. Therefore, students can use intermediate testing to understand their own retrieval strategies,

monitor their learning, and allocate their study time more efficiently (Karpicke, 2009). The important point is that testing provides a diverse array of benefits for remembering a set of material. Although individual differences research has not examined direct effects of retrieval, one previous study has investigated the indirect effects of retrieval and this work will be reviewed in the following section.

Individual Differences and the Effects of Testing

As discussed at the beginning of this manuscript, an individual differences approach can be a useful tool for comparing competing theories of memory as well as elucidating the component processes underlying memory ability (Underwood, 1978). Recently, we have demonstrated that low working memory participants can capitalize on free recall testing to minimize the accrual of proactive interference across semantically related word lists thus bringing their recall performance to levels resembling that of high working memory participants (Brewer, Unsworth, & Spillers, 2010). Thus, at the very least, individual differences manifest themselves in an indirect effect of retrieval (i.e., interference reduction). With regards to direct effects of testing, Chan (2009) collected data from a working memory task (operation span) but failed to find any relationships with the testing effect. One problem with correlating performance between testing effects and external correlates such as working memory is that difference scores can be notoriously unreliable. Also, multiple measures of a construct should be collected whenever possible to avoid idiosyncratic task-related relationships. Therefore, it is an open question as to when and how working memory will be related to the magnitude of the testing effect (viz. direct versus indirect benefits). Furthermore, there are multiple cognitive constructs in addition to working memory that may be related to the testing effect. In order to fully account for the variation in direct effects of retrieval from long-term memory, multiple

constructs should be examined. To date, no published study has focused explicitly on examining individual differences in working memory, attention control, episodic memory, intelligence and the direct effects of testing memory.

Working memory is the system responsible for actively maintaining information in the face of distraction (Baddeley, 2007; Jonides, Lacey, & Nee, 2005). Performance on working memory tasks is related to various higher order cognitive abilities including reading comprehension, standardized achievement test scores, reasoning ability, and intelligence (see Engle & Kane, 2004 for a review). Working memory capacity and higher order cognition are both reliant on the flexible control of attention and memory processes (Unsworth & Engle, 2007). As reviewed earlier, research has suggested that effortful retrieval leads to the largest testing effects. To the degree that attention control is necessary for effortful retrieval (Kane & Engle, 2000), performance on basic attention tasks may be related to the magnitude of the testing effect. Therefore, the strategic regulation of attention may be an important factor for understanding individual differences in the testing effect (Dudukovic, DuBrow, & Wagner, 2009). Given that memory retrieval is a basic component of testing effect paradigms, it seems most likely that tasks tapping episodic memory abilities will be closely related to the size of the testing effect at an individual differences level. Both attention control and episodic retrieval processes underlie the relation between working memory capacity and general-fluid intelligence (gF; Unsworth & Spillers, in press). As such, gF may share important variation with the magnitude of the testing effect to the degree that gF broadly represents fluid reasoning and domain general cognitive control abilities. In the current study, the relation between the testing effect and these concomitant variables will be investigated.

Research has suggested that retrieval is an important determinant of both remembering and forgetting but one critical question remains unanswered: How would performance on basic attention, memory, and intelligence tasks predict individual differences in the testing effect? Primarily, there are three hypotheses that are derived from taking an individual differences approach to exploring the effects of retrieval on long-term memory.

1. Testing provides general benefits across the ability range (invariance across participants).
2. Testing allows the rich to get richer (high ability students are helped more than low ability students).
3. Testing homologizes memory across the ability range (low ability students behave like high ability students).

In these hypotheses, "ability" refers to performance on various working memory, attention, episodic memory, and intelligence tasks.

The Current Study

The current study employed a large scale individual differences approach with multiple measures of working memory capacity (operation, symmetry, and reading span), attention control (antisaccade, arrow flankers, and psychomotor vigilance), episodic memory (delayed free recall, cued recall, and gender and picture source monitoring), and intelligence (Raven, number series, and letter series). Composite scores were drawn out of these measures and were used to predict performance on a paired-associate testing task. Specifically, in the testing task, participants studied multiple cue-target pairs and then took a test over half of the pairs and restudied the other half (e.g., Carpenter et al., 2006). Having multiple measures of each construct allowed us to compute composite scores with better psychometric properties. Performance on the testing task was analyzed, relations with external correlates were examined,

and analysis of covariance methods were implemented to test the three hypotheses given earlier and to investigate individual differences in the testing effect.

METHOD

Participants

One hundred and seven undergraduate students from the University of Georgia volunteered in exchange for credit toward a research appreciation requirement. Each participant was tested on a computerized battery of tasks that measured the testing effect in paired-associate learning, working memory capacity, attention control, episodic memory, and intelligence. Each participant was tested in two sessions lasting approximately 2 hours each.

Materials and Procedure

After signing informed consent, all participants completed Operation Span, Symmetry Span, Reading Span, Cued Recall, Paired-Associate Testing Task (intermediate testing), and Number Series in Session 1. In Session 2, all participants completed a Picture-Source Recognition, Raven Advance Progressive Matrices, Flanker Task, Psychomotor Vigilance Task, Delayed Free Recall Task, Letter Sets, Paired-Associate Testing Task (final testing), and an Antisaccade task. All tasks were administered in the order listed above and the two testing sessions were separated by a 24 hour delay.

Tasks

Paired-Associate Testing Task. The parameters of this task closely resembled those used by Carpenter, Pashler, and Vul (2006). In this task participants encoded 40 word pairs. Subsequent to the study phase, participants restudied 20 of the cue-target pairs and then took a cued-recall test over the other 20 pairs. In the initial test participants were presented with the cue word and instructed to type the target word that it was originally paired with during the encoding

phase. In the second experimental session (separated by 24 hours) participants were tested over all 40 cue-target pairs. The dependent variable was the proportion of originally tested and restudied cue-target pairs correctly recalled on the final test.

Working Memory Tasks.

Operation Span (Ospan). Participants solved a series of math operations while trying to remember a set of unrelated letters. Participants were required to solve a math operation and after solving the operation they were presented with a letter for 1 s. Immediately after the letter was presented the next operation was presented. At recall, letters from the current set were recalled in the correct order by clicking on the appropriate letters. For all of the span measures, items were scored if the item is correct and in the correct serial position. The dependent variable is the number of correct items recalled in the correct serial position.

Symmetry Span (Symspan). Participants were required to recall sequences of red squares within a matrix while performing a symmetry-judgment task. In the symmetry-judgment task participants were shown an 8 x 8 matrix with some squares filled in black. Participants decided whether the design was symmetrical about its vertical axis. The pattern was symmetrical half of the time. Immediately after determining whether the pattern was symmetrical, participants were presented with a 4 x 4 matrix with one of the cells filled in red for 650 ms. At recall, participants recalled the sequence of red-square locations in the preceding displays, in the order they appeared by clicking on the cells of an empty matrix. The same scoring procedure as Ospan was used.

Reading Span (Rspan). Participants were required to read sentences while trying to remember a set of unrelated letters. Participants read a sentence and determined whether the sentence made sense or not. Half of the sentences made sense while the other half did not.

Nonsense sentences were made by simply changing one word from an otherwise normal sentence. After participants gave their response they were presented with a letter for 1 s. At recall, letters from the current set were recalled in the correct order by clicking on the appropriate letters. The same scoring procedure as Ospan was used.

Attention Control Tasks.

Antisaccade. In this task (Kane, Bleckley, Conway, & Engle, 2001) participants were instructed to stare at a fixation point which is onscreen for a variable amount of time (200-2200 ms). A flashing white "=" was then flashed either to the left or right of fixation (11.33° of visual angle) for 100 ms. This cue was followed by the target stimulus (a B, P, or R) onscreen for 100 ms. The target was followed by masking stimuli (an H for 50 ms and an 8 which remains onscreen until a response is given). The participants' task was to identify the target letter by pressing a key for B, P, or R (the keys 1, 2, or 3) as quickly and accurately as possible. In the prosaccade condition the flashing cue (=) and the target appeared in the same location. In the antisaccade condition the target appeared in the opposite location as the flashing cue. Participants received, in order, 10 practice trials to learn the response mapping, 15 trials of the prosaccade condition, and 60 trials of the antisaccade condition. The dependent variable was the proportion of errors on the antisaccade trials.

Arrow Flankers. Participants were presented with a fixation point for 400 ms. This was followed by an arrow directly above the fixation point for 1700 ms. The participants' task was to indicate the direction the arrow was pointing (pressing the F for left pointing arrows or pressing J for right pointing arrows) as quickly and accurately as possible. On 50 neutral trials the arrow was flanked by two horizontal lines on each side. On 50 congruent trials the arrow was flanked by two arrows pointing in the same direction as the target arrow on each side. Finally, on 50

incongruent trials the target arrow was flanked by two arrows pointing in the opposite direction as the target arrow on each side. All trial types were randomly intermixed. The dependent variable was the reaction time difference between incongruent and congruent trials.

Psychomotor Vigilance Task (PVT). Participants were presented with a row of zeros on screen and after a variable amount of time the zeros began to count up in 1 ms intervals from 0 ms. The participants' task was to press the spacebar as quickly as possible once the numbers started counting up. After pressing the spacebar the RT was left on screen for 1 s to provide feedback to the participants. Interstimulus intervals were randomly distributed and ranged from 1 to 10s. The entire task lasted for 10 minutes for each individual (roughly 75 total trials). The dependent variable is the average reaction time from the slowest 20% of trials (Dinges & Powell, 1985).

Episodic Memory Tasks.

Delayed Free Recall Unrelated Words. Participants attempted to recall 6 lists of 10 words each. All words were common nouns that were presented for 1s each. After list presentation, participants had a distractor task for 16s in which a three-digit number appeared for 2s and then they wrote the digits in ascending order on a separate piece of paper. After the distractor task participants typed as many words as they could remember from the current list in any order they wished. Participants had 45s for recall. A participant's score was the total number of items recalled correctly.

Cued Recall. In this task, participants were given 3 lists of 10 words pairs each. All words were common nouns and the word pairs were presented vertically for 2 s each. Participants were told that the cue would always be the word on top and the target would be on bottom. After the presentation of the last word, participants saw the cue word and ??? in place of

the target word. Participants were instructed to type in the target word from the current list that matched cue and then to press ENTER to indicate their response. The cues were randomly mixed so that the corresponding target words were not recalled in the same order as they were presented. Participants had 5s to type in the corresponding word. This same procedure was done for all three lists. A participant's score was the proportion of items recalled correctly.

Picture Source-Recognition. Participants were presented with a picture (30 total pictures) in one of four different quadrants onscreen for 1 s. Participants were explicitly instructed at encoding to pay attention to both the picture as well as the quadrant it was located in. At test, participants were presented with 30 old and 30 new pictures in the center of the screen. Participants indicated if the picture was new or old and, if old, what quadrant it was originally presented in via key press. Participants had 5 s to press the appropriate key to enter their response. A participant's score was the proportion correct.

Gender Source Recognition. Participants heard words (30 total words) in either a male or a female voice. Participants were explicitly instructed to pay attention to both the word as well as the voice the word was spoken in. At test participants were presented with 30 old and 30 new words and were required to indicate if the word was new or old and, if old, what voice it was spoken in via key press. Participants had 5 s to press the appropriate key to enter their response. A participant's score was the proportion of correct responses.

Intelligence Tasks.

Raven Advanced Progressive Matrices. The Raven consisted of 18 items presented in escalating degree of difficulty. Each item consisted of a display of 3 x 3 matrices of geometric patterns with the bottom right pattern missing. The task required participants to select, among eight alternatives, the one that correctly completed the overall series of patterns. Participants had

10 minutes to complete the 18 odd-numbered items. A participant's score was the total number of correct solutions.

Number Series. In this task, participants saw a series of numbers and determined what the next number in the sequence should be (Thurstone, 1962). That is, the series followed some unstated rule which participants were required to figure out in order to determine which the next number in the series should be. Participants selected their answer out of five possible numbers that were presented. Following five practice items, participants had 4.5 minutes to complete 15 test items. A participant's score was the total number of items solved correctly.

Letter Sets. On each problem, participants saw five sets of letters containing four letters each. Participants were instructed to find the rule that applied to four of the five letter sets, and then indicate the letter set that violated the rule. Participants had 5 minutes to complete 20 items, with their total correct used as the dependent variable.

RESULTS

Descriptive statistics and correlations for all measures can be found in Tables 1 and 2, respectively. In line with previous research, z-score composites were made for the WMC (ospan, symspan, rspan), AC (antisaccade, arrow flankers, pvt), episodic (delayed free recall, cued recall, picture source), and intelligence (raven, number series, letter series) constructs. Replicating much previous research, these z-scores were interrelated (see Table 3).

The Testing Effect.

The testing effect in the paired-associate testing task was replicated in the current study, $F(1,106) = 25.16$, $p < .001$, $\eta^2_p = .19$. That is, more targets were successfully recalled if they had previously been tested (M=.51) as compared with restudied (M=.44). Additionally, there was a great deal of variability in participants' susceptibility to the effects of testing.

Individual Differences in the Testing Effect.

Figure 1 depicts a frequency histogram of each participant's difference between their performance in testing and nontesting conditions. This difference score was calculated by subtracting the mean proportion of items successfully recalled in the nontested pairs from the mean proportion recalled in the tested pairs. Thus, difference scores greater than 0 indicate a positive testing effect, scores equal to 0 indicate no effect, and scores less than 0 indicate a negative testing effect. The astute reader will notice that positive benefits of testing did not hold for all participants in this experiment. Approximately 71 participants demonstrated a positive testing effect, 13 demonstrated no testing effect, and 23 demonstrated a negative testing effect.

In order to further distinguish these groups, several measures of performance were derived from the initial and final testing phases of the paired-associate testing task. The average percentage of words recalled in the initial testing phase was 47%. This measure was not related to the difference score which reflected the effects of testing, $r = .004$. Therefore, the total number of words recalled in the initial testing phase did not determine the magnitude of the testing effect in the subsequent test. Overall recall performance on final testing task was defined as mean performance pooling over cue-target pairs in the testing (M = .51) and nontesting (M = .44) conditions. There was a significant relation between the overall proportion correctly recalled and magnitude of the difference score, $r = -.28$, $p < .05$. This correlation seemed to be primarily driven by the large correlation between proportion correct for the nontesting pairs and the difference score, $r = -.50$ (the correlation between testing pairs and the difference score was $r = -.01$). This correlation is consistent with the notion that participants with worse memory performance in the paired-associate testing task tended to show the most positive effects of initial testing. That is, poor performers on the criterion task tend to have the biggest positive benefits

from testing but it remains to be seen whether other external correlates are related to the magnitude of the testing effect.

Relations with External Measures.

To investigate the relationships between the paired-associate testing effect and external measures of WMC, AC, episodic memory, and gF all z-scores were simultaneously entered into a repeated measures analysis of covariance. The repeated measures factor was composed of the nontesting and testing means in the paired associate testing task. The broad correlations amongst these measures are found in Table 2 and the correlations amongst the z-composites are found in Table 3. There was no significant interaction between either WMC or AC and the magnitude of the testing effect. There was, however, a significant interaction between the episodic memory composite and the magnitude of the testing effect, $F(1,102) = 5.14$, $p < .05$, $\eta^2_p = .05$. Additionally, there was a significant interaction between the magnitude of the testing effect and gF, $F(1,102) = 4.94$, $p < .05$, $\eta^2_p = .05$. [1]

Careful examination of Figure 2a shows that individuals with low episodic memory z-scores typically exhibited bigger testing effects than participants with higher episodic memory scores. To further investigate this effect, participants falling in the upper and lower quartiles of the distribution of episodic memory scores were selected. As can be seen in Figure 2b, participants who were in the lower quartile of episodic memory performance exhibited significantly larger testing effects than participants in the upper quartile, $F(1,51) = 6.01$, $p < .05$, $\eta^2_p = .11$. Follow up t-tests confirmed that participants falling in the lower quartile of the episodic memory scores exhibited a significant testing effect whereas participants in the upper quartile did not exhibit a significant effect, $t(27) = 5.41$, $p < .01$ and $t(26) = 1.61$, *ns.* Similar effects were found when examining the relation between gF and the paired-associate testing

effect (see Figures 3a and 3b). More specifically, participants in the lower quartile of gF scores showed the largest testing effects, $F(1,52) = 6.54$, $p < .05$, $\eta^2_p = .11$. Follow up t-tests confirmed that participants falling in the lower quartile of gF scores exhibited a significant testing effect whereas participants in the upper quartile did not exhibit a significant effect, $t(27) = 4.97$, $p < .01$ and $t(27) = 1.39$, *ns.* Thus, of the four external cognitive correlates examined in the current study, only episodic abilities and general-fluid intelligence correlated with the paired-associate testing effect. Furthermore, these correlations were both negative indicating that low ability students benefit more from paired-associate testing than high ability students.

Discussion

The results from this study speak to the nature of individual differences in the direct effects of intermediate paired-associate testing in several important ways. First, the testing effect reported by Carpenter and colleagues (2006) was replicated. This is not surprising given that nearly identical materials and procedures were implemented in the current study. Second, there was variability in the magnitude of the testing effect with some participants demonstrating large effects, some demonstrating no effects, and others demonstrating negative effects. Therefore, much individual variability existed in the magnitude of the effects of initial testing over mere restudying. Third, performance on the criterion paired-associate testing task was related to the magnitude of the testing effect. Specifically, participants who generally exhibited worse overall memory in the task had larger positive testing effects. Fourth, external measures of WMC, AC, episodic memory, and gF were interrelated (replicating previous research; Unsworth & Spillers, in press), related to performance on the criterion paired-associate testing task, but only the episodic memory and gF constructs were related to the magnitude of the testing effect. Collectively, the results from the current study inform extant research on the beneficial effects of

retrieval from long-term memory and speak to the underlying theoretical mechanisms that support the testing effect.

As described previously, there are multiple ways that retrieval from long-term memory enhances subsequent memory. In the current study, it was hypothesized that initial paired-associate testing would lead to more accessible memories for future retrieval attempts. The retrieval-effort hypothesis assumes that effortful control over memory search leads to more durable representation in memory by the creation of a variety of retrieval routes. In the current study, the cue-target pairs had intermediate preexperimental associative values so retrieval may not has been as effortful as if the pairs were completely unrelated. Perhaps the degree of effortful control that is exerted during initial testing is related to the magnitude of the testing effect (for a discussion of this issue see Carpenter, in press). An important direction for future individual differences research will be to investigate other paired-associate testing tasks (e.g., no preexperimental association between the cue-target pairs) to examine variability in the magnitude of the effect. These types of investigations may have a great deal of explanatory power by potentially shifting the entire distribution of participants toward exhibiting positive benefits of testing, as opposed to the current study where a moderate percentage (34%) failed to show a positive effect. Clearly, an important direction for both applied and experimental memory researchers is to investigate individual differences in the testing effect at the participant level.

Along these lines, the participants with impoverished performance on the paired-associate testing task were generally those participants who demonstrated the largest positive testing effects. This relationship was primarily driven by performance for the nontested cue-target pairs. This result indicates that individuals with poor memory for the nontested pairs in

the criterion task had the biggest effects of testing. Although counterintuitive, this result is consistent with the notion that tested material is more resistant to the effects of forgetting over time (Karpicke & Roediger, 2007). Perhaps some participants implemented testing-based learning strategies for the nontested pairs as well as the tested pairs during initial presentation. On the one hand, these participants would not have exhibited as steep of a forgetting function for the nontested pairs because they encoded those items using more efficient strategies (i.e., testing, elaboration, imagery, etc.). On the other hand, participants who passively encoded the nontested pairs may have had the most precipitous drops in their levels of recall for those pairs. In this manner, forgetting of nontested information leads to bigger testing effects because the tested pairs are more durable to forgetting. Future investigations of individual differences in the testing effect should pay close attention to the criterion task and it's relation with the overall benefits of testing. To gain explanatory leverage on the variability in the testing effect, external cognitive correlates also must be measured and assessed.

In the current study, multiple measures of working memory, attention control, episodic memory, and higher-order intelligence were collected. Neither the WMC nor AC constructs were reliably related to the magnitude of the testing effect (see Chan, 2009 for similar findings with a single measure of working memory). These results stand in opposition to previously reported results on the relation between WMC and the indirect effects of retrieval from long-term memory. Brewer et al (2010) found that students with poor working memory abilities can benefit from free recall testing to reduce proactive interference which accrues across multiple study-test trials on semantically related word lists (see also Szpunar, McDermott, & Roediger, 2008). Taken together, the results from the current study and Brewer et al (2010) highlight the notion that working memory is related to the testing effect in only certain circumstances. As

described earlier, Roediger and Karpicke (2006) proposed that testing leads to better memory performance several different means. Direct effects of testing are those that influence the accessibility of the specifically tested information by making it more durable, stronger, and distinctive. Indirect (or mediated) effects of testing are those that influence participants metacognitive monitoring and control processes such as study-time allocation, source monitoring, and cue-driven retrieval strategies (Karpicke, 2009; Szpunar et al., 2008). In Brewer et al (2010), low WMC students used testing to sharpen their cue-driven retrieval strategies and we proposed that this testing effect was indirect in nature (for a description of variation in WMC from cue driven search see Unsworth, 2007). The current results demonstrate that direct effects of testing, at least in the paired-associate testing paradigm, are not related to WMC or AC (see also Chan, 2009). Thus, it remains an open question as to when external measures of attentional control or working memory will be related to the direct effects of testing when a greater deal of effort is needed during intermediate retrieval attempts.

Episodic memory ability was, however, related to the magnitude of the paired-associate testing effect. This relationship was primarily driven by low episodic ability students showing bigger benefits from testing than high ability students. This result is consistent with the previous finding that poor performers on the criterion task (paired-associate testing) showed the biggest benefits from testing. This difference could arise for a number of reasons. For instance, participants with more efficient episodic memory processes may not benefit as much from testing because they elaborately encode information leading to multiple retrieval routes whereas participants with poor episodic memory ability need intermediate retrieval to build these routes (Bjork, 1975; Carpenter, in press). By this hypothesis, high ability participants are at a functional ceiling (although performance was well below 100%). Another possibility is that

participants who were in the lower in episodic memory performance range may have learned more efficient retrieval strategies during the first testing session (Cokely, Kelley, & Gilchrist, 2006). These retrieval strategies may have benefited the retrieval of both the tested and restudied pairs during the final criterion test. A final possibility is that there may be another factor underlying the relation between episodic memory ability and the testing effect. Although, it is difficult to imagine what other factor may mediate this relation because the two constructs are so similar. Thus, future work is needed to tease apart these competing explanations for the current results. Nevertheless, students with poor episodic memory abilities benefit more from testing on material in paired-associate tasks and this finding is clearly of great importance for applied and educational psychologists.

Perhaps the most compelling finding reported in the present work was that measures of higher-order intelligence were related to the magnitude of the paired-associate testing effect when controlling for other variation in other external measures of working memory, attention, and episodic memory. Currently, it is not yet known why gF is correlated with the testing effect but there are several potential hypotheses to be tested in future research. Perhaps this tantalizing correlation is driven by a general g-factor that extends across a variety of mental abilities including the testing effect (Jensen, 1998). Alternatively, there may exist some component tapped by the common variance amongst Raven, number series, and letter series that is responsible for the correlation. This component may be related to metacognitive monitoring and control processes. With regards to the hypotheses proposed in the introduction, the results from the current study are most consistent with the idea that testing homologizes performance across the ability range; although, intermediate retrieval did not completely equate students with low and high gF. The intelligence-testing effect relationship demonstrated herein should resonate

with researchers who are actively implementing testing procedures in classrooms and other applied settings (McDainel, Roediger, & McDermott, 2007). The current results indicate that future work in applied psychological settings begin examining and accounting for individual differences in the classroom environment.

Conclusion

The primary goal of the current research was to implement a large scale individual differences study of the testing effect and uncover central relations with higher-order cognition. It is an important, and unanswered question whether the testing effect extends to all students in the same manner, helps high ability students more than low ability students, or helps low ability students more than high ability students. With regards to Neisser and colleague's (1996) quote at the head of this report, the current research points to a specific role of the testing effect in ameliorating the correlation between intelligence and test scores in a group of college students. However, the correlation has not yet been fully eliminated. This research clearly demonstrates an important relation between the testing effect and episodic memory and intelligence abilities. Clearly much more research of this nature is needed.

References

Abbott, E.E. (1909). On the analysis of the factors of recall in the learning process. Psychological Monographs, 11, 159–177.

Baddeley, A. D. (in press). Working memory, thought and action. Oxford, UK: Oxford University Press.

Bjork, R.A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R.L. Solso (Ed.), Information processing and cognition: The Loyola Symposium (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R.A. (1988). Retrieval practice and the maintenance of knowledge. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), Practical aspects of memory: Current research and issues (Vol. 1, pp. 396–401). New York: Wiley.

Brewer, G. A., Unsworth, N., & Spillers, G. J. (2010). Working memory, interference, and the testing effect. *Manuscript Currently Under Review.*

Calkins, M. W. (1894). Association: I. *Psychological Review, I.* 476- 483.

Carpenter, S. K. (in press). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Carpenter, S.K., & DeLosh, E.L. (2005). Application of the testing and spacing effects to name learning. Applied Cognitive Psychology, 19, 619–636.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268-276.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? Psychonomic Bulletin & Review, 13, 826-830.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. Memory & Cognition, 20, 633–642.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. Journal of Memory and Language, 61, 153-170.

Chan, J.C.K., & McDermott, K.B.(2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 431-437.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 215–235.*

Dinges, D. F., & Powell, J.W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers*, 17, 652-655.

Dudukovic, N. M., DuBrow, S., & Wagner, A. D. (2009). Attention during memory retrieval enhances future remembering. *Memory & Cognition, 37,* 953-961.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 145–199). New York: Elsevier.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62,* 145-154.

Glover, J.A. (1989). The ‘‘testing’’ phenomenon: Not gone but nearly forgotten. Journal of *Educational Psychology, 81,* 392–399.

Izawa, C. (1967). Function of test trials in paried-associate learning. *Psychological Reports*, *75*, 194-209.

Izawa, C. (1969). Comparison of reinforcement and test trials in paired associate learning. *Journal of Experimental Psychology*, 81, 600-603.

Jensen, A. (1998). The G factor: The science of mental ability. Greenwood Publishing Group.

Jonides, J., Lacey, S.C., and Nee, D.E. (2005). Processes of working memory in mind and brain. *Current Directions in Psychological Science, 14*, 2-5.

Kane, M.J., Bleckley, M.K., Conway, A.R.A., & Engle, R.W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130*, 169-183.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26,* 336-358.

Kang, S.H.K., McDermott, K.B. & Roediger, H.L. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *The European Journal of Cognitive Psychology*, *19*, 528-558.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469-486.

McDaniel, M.A. (2007). Transfer. In H.L. Roediger, III, Y. Dudai, & S.M. Fitzpatrick (Eds.), The science of learning and memory: Concepts. Oxford, England: Oxford University Press.

McDaniel, M.A., & Masson, M.E.J. (1985). Altering memory representations through retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 371–385.

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, **14**, 200-206.

Murdock, B. B. (1974). Human memory: Theory and data. Hillsdale, NJ: Erlbaum.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol.26, pp. 125–173). New York: Academic Press.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437-447.

Raven, J.C., Raven, J.E., & Court, J.H. (1998). *Progressive Matrices*. Oxford, England: Oxford Psychologists Press.

Roediger, H.L., Agarwal, P.K., Kang, S.H.K, & Marsh, E.J. (in press). Benefits of testing memory: Best practices and boundary conditions. In G.M. Davies & D.B. Wright (Eds.), New frontiers in applied memory. Brighton, U.K.: Psychology Press.

Roediger, H.L. & Karpicke, J.D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210.

Thurstone, T.G. (1962). *Primary Mental Abilities.* Chicago, Science Research Associates.

Underwood, B.J. (1975). Individual differences as a crucible in theory construction. *American Psychologist, 30,* 128-134.

Unsworth, N. (2007). Individual differences in working memory capacity and episodic retrieval: Examining the dynamics of delayed and continuous distractor free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 1020-1034.

Unsworth, N., & Engle, R.W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. Psychological Review, 114, 104-132.

Unsworth, N. & Spillers, G. J. (in press). Working memory capacity: Attention, memory, or both? A direct test of the dual-component model. *Journal of Memory & Language*.

Footnote

1. Although difference scores can be unreliable, we report here the correlations between the difference score (testing – nontesting) and the z-scores for episodic memory (r = -.29) and gF (r = -.28). Both correlations were statistically significant, whereas the correlations with WMC (r = -.12) and AC (.08) were nonsignificant.

Table 1. Descriptive statistics for all of the measures.

Measure	Mean	S.D.	Skew	Kurtosis
PAT Testing	0.51	0.22	0.20	-0.73
PAT Nontesting	0.45	0.26	0.43	-0.78
PAT Diff	0.06	0.13	-0.29	-0.27
Ospan	60.51	12.49	-1.56	2.87
Sym	30.12	7.66	-0.51	-0.46
Rspan	59.98	11.99	-1.33	1.82
Anti	0.50	0.14	-0.28	-0.31
Vigilance	504.94	152.24	2.15	5.85
Flanker	109.54	60.23	1.62	3.47
Gsource	0.62	0.12	-0.15	0.17
Psource	0.81	0.11	-1.37	3.01
CR	0.49	0.23	0.07	-0.87
DFR	0.54	0.19	0.08	-1.90
Nseries	0.72	0.15	-0.97	2.17
Lseries	0.69	0.17	-1.25	1.12
Raven	0.59	0.13	-0.52	0.52

Note: PAT Testing = paired-associate testing condition; PAT Nontesting = paired-associate nontesting condition; PAT Diff = Difference Score; Ospan = operation span; Sym = symmetry span; Rspan = reading span; Anti = antisaccade; Vigilance = slowest reaction times; Flanker = flanker effect; Gsource = gender source; Psource = picture source; CR = cued recall; DFR = delayed free recall; Nseries = number series; Lseries = letter series; Raven = Raven advanced progressive matrices.

Table 2. Correlations for all of the measures.

	PATTest	PATNon	PATDiff	Ospan	Sym	Rspan	Anti	Vigilance	Flanker	Gsource	Psource	CR	DFR	Nseries	Lseries	Raven
PATTest	1.00															
PATNon	0.86	1.00														
PATDiff	-0.01	-0.50	1.00													
Ospan	0.24	0.26	-0.10	1.00												
Sym	0.14	0.23	-0.22	0.38	1.00											
Rspan	0.41	0.34	0.04	0.68	0.33	1.00										
Anti	-0.10	-0.11	0.04	-0.07	-0.21	0.00	1.00									
Vigilance	-0.27	-0.28	0.09	-0.16	-0.15	-0.09	0.39	1.00								
Flanker	-0.12	-0.13	0.06	-0.22	-0.13	-0.18	0.13	0.10	1.00							
Gsource	0.23	0.27	-0.15	0.16	0.19	0.12	-0.20	-0.24	-0.08	1.00						
Psource	0.42	0.48	-0.23	0.34	0.21	0.47	-0.08	-0.11	-0.22	0.33	1.00					
CR	0.51	0.51	-0.14	0.17	0.34	0.26	-0.05	-0.04	-0.04	0.33	0.27	1.00				
DFR	0.45	0.54	-0.29	0.40	0.34	0.37	-0.12	-0.35	-0.13	0.24	0.30	0.58	1.00			
Nseries	0.16	0.23	-0.19	0.18	0.27	0.09	-0.12	-0.15	-0.10	0.17	0.27	0.11	0.13	1.00		
Lseries	0.23	0.35	-0.31	0.13	0.26	0.05	-0.19	-0.46	-0.18	0.18	0.20	0.13	0.28	0.46	1.00	
Raven	0.19	0.26	-0.18	0.10	0.35	0.07	-0.21	-0.37	-0.18	0.29	0.17	0.16	0.19	0.41	0.52	1.00

Note: PATTest = paired-associate testing condition; PATNon = paired-associate nontesting condition; PATDiff = Difference Score; Ospan = operation span; Sym = symmetry span; Rspan = reading span; Anti = antisaccade; Vigilance = slowest reaction times; Flanker = flanker effect; Gsource = gender source; Psource = picture source; CR = cued recall; DFR = delayed free recall; Nseries = number series; Lseries = letter series; Raven = Raven advanced progressive matrices.

Table 3. Correlations for the composite scores.

	WMC	AC	EPI	gF
WMC	1.00			
AC	-0.23	1.00		
EPI	0.49	-0.26	1.00	
gF	0.26	-0.36	0.33	1.00

Note: PATDiff = testing proportion correct – nontesting proportion correct; WMC = z-composite for three complex-span tasks; AC = z-composite for three attention-control tasks; EPI = z-composite for four episodic memory tasks; gF = z-composite for three intelligence tasks.

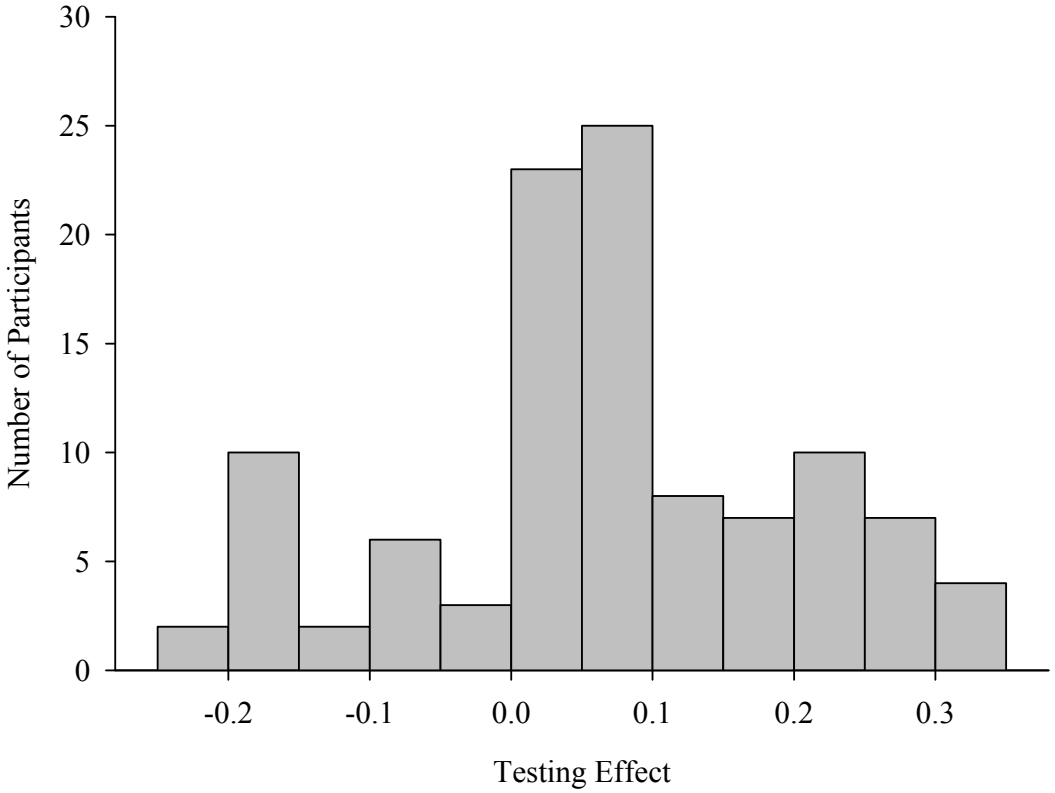
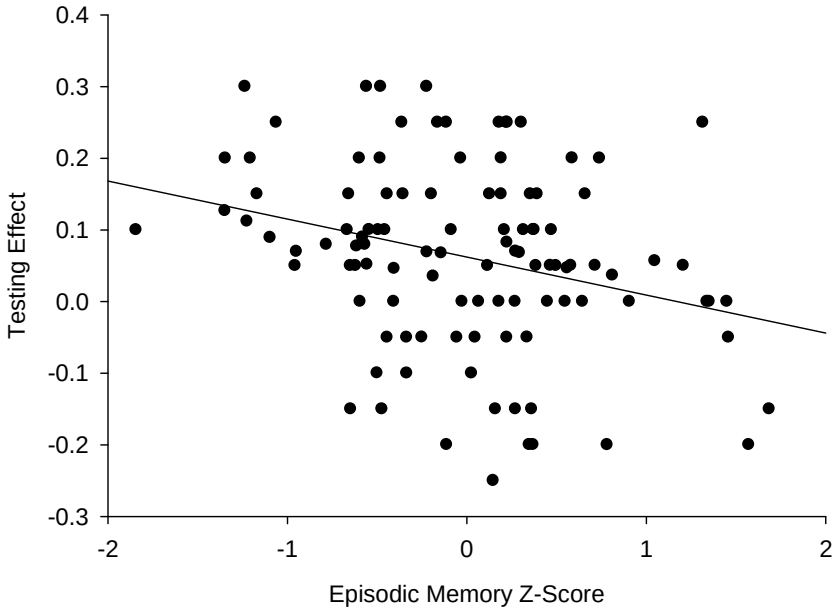


Figure 1. A Histogram showing the distribution of scores reflecting the testing effect.

(a)



(b)

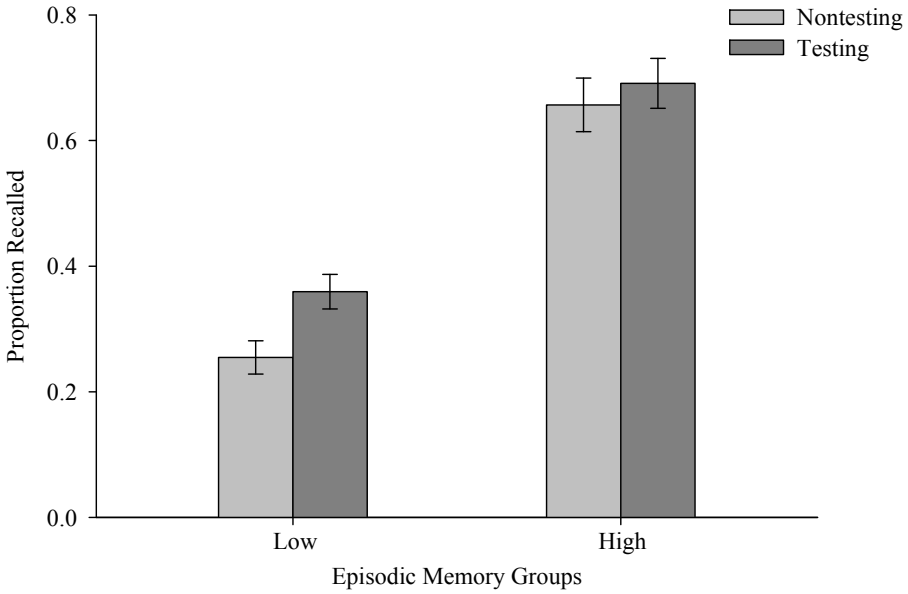
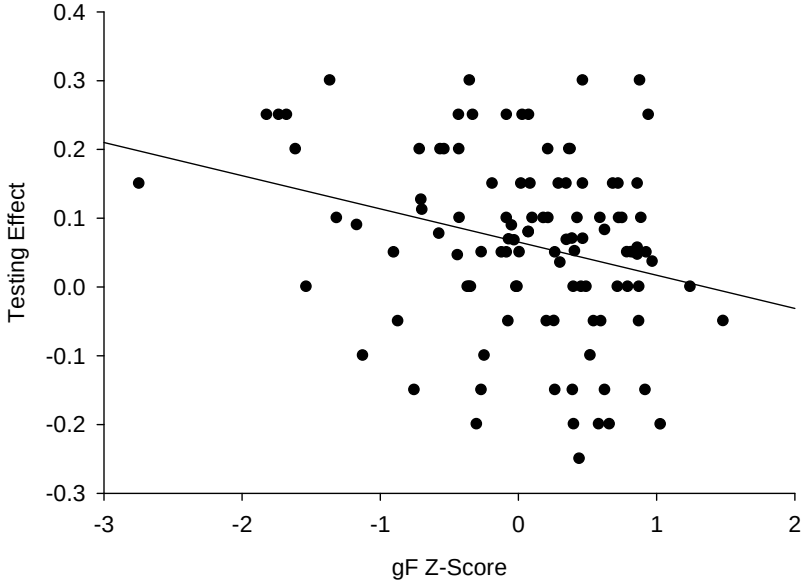


Figure 2. A scatterplot (a) and bar graph (b) showing the relation between episodic memory abilities and the magnitude of the testing effect.

(a)



(b)

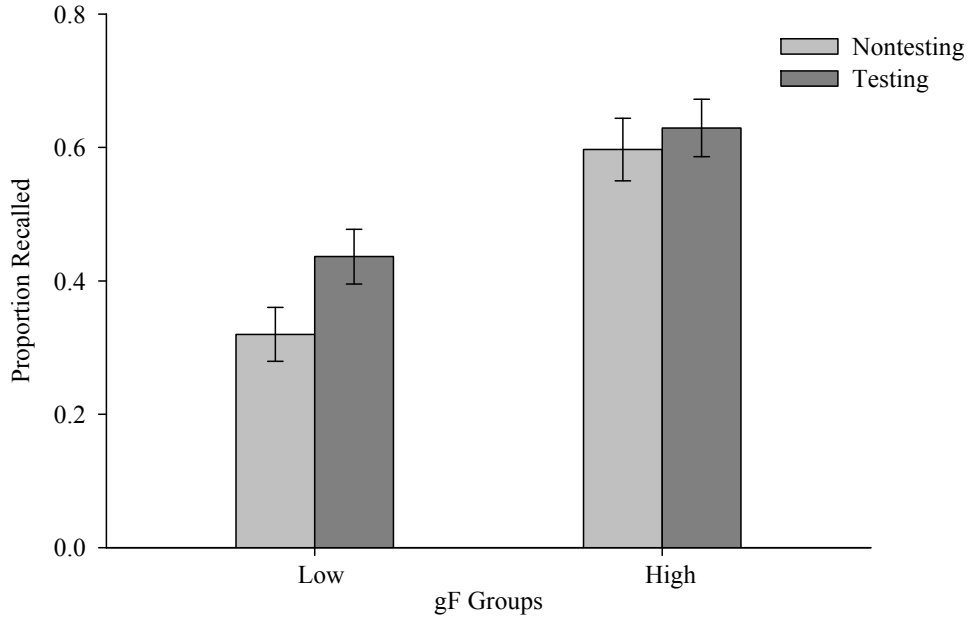


Figure 3. A scatterplot (a) and bar graph (b) showing the relation between general-fluid intelligence and the magnitude of the testing effect.